REMARKS

I. Status of the Claims

Claims 1-20 were originally filed. Claims 3-20 have been withdrawn from consideration as the result of a restriction requirement. Claims 1 and 2 were pending under examination.

Upon entry of the present amendment, claims 3-20 are canceled. The word "a" is deleted from the phrase "conferring mitoxantrone resistance to a S1-M1-80 human colon carcinoma cells when expressed in the cells" in claim 1 to correct a grammatical error. The recitation of SEQ ID NO:4 is deleted from claims 1 and 2. New claim 21 is added, support for which is found throughout the specification and in original claims. No new matter is introduced.

II. Claim Rejections

A. 35 U.S.C. §112, First Paragraph: Written Description

Claim 1 was rejected under 35 U.S.C. §112, first paragraph, as the Examiner alleged that the specification does not provide adequate description of the claimed subject matter so as to reasonably convey to one of skill in the art that the inventors, at the time of filing, had possession of the claimed invention. Specifically, the Examiner stated that a protein capable of "specifically binding to polyclonal antibodies which specifically bind to a member of the group of proteins depicted in SEQ ID NO:2 or SEQ ID NO:4" is not adequately described. Applicants respectfully traverse the rejection.

Possession of claimed invention may be shown by a variety of descriptive means, including words, structure, figures, diagrams, and formulas. MPEP §2163 I. Case law provides more specific guidance in setting the standard for written description.

Claim 1 as amended is directed to an isolated ATP-binding cassette protein having the following properties: (i) conferring mitoxantrone resistance to S1-M1-80 human colon carcinoma cells when expressed in the cells; (ii) specifically binding to polyclonal antibodies which specifically recognize a protein having the amino acid sequence depicted in

SEQ ID NO:2; and (iii) having a molecular weight between about 70 kDa and about 75 kDa. This claim fully complies with the requirements for written description of a chemical genus as set forth in *University of California v. Eli Lilly & Co.*, 43 USPQ2d 1398 (Fed. Cir. 1997). As described by the Federal Circuit in *Lilly*, "[a] description of a genus of cDNAs may be achieved by means of . . . a recitation of structural features common to the members of the genus . . . ." *Lilly*, 43 USPQ2d at 1406. Furthermore, the court in *Fiers v. Revel* stated that an adequate written description "requires a precise definition, such as by structure, formula, chemical name, or physical properties." *Fiers*, 25 USPQ2d 1601, 1606 (Fed. Cir. 1993).

On the other hand, proper description of functional features of a claimed invention can play an important role in satisfying the written description requirement. The Federal Circuit recently stated that "*Lilly* did not hold that all functional descriptions of genetic material necessarily fail as a matter of law to meet the written description requirement; rather, the requirement may be satisfied if in the knowledge of the art the disclosed function is sufficiently correlated to a particular, known structure." *Amgen Inc. v. Hoechst Marion Roussel Inc.*, 65 USPQ2d 1385, 1398 (Fed. Cir. 2003).

With regard to the claimed polypeptides, claim 1 sets forth both functional features, *e.g.*, conferring mitoxantrone resistance to S1-M1-80 human colon carcinoma cells when expressed in the cells, and structural features, *e.g.*, specifically binding to polyclonal antibodies which specifically bind to a protein having the amino acid sequence depicted in SEQ ID NO:2.

The ability for a polypeptide to bind specifically to polyclonal antibodies that specifically recognize a reference amino acid sequence, is a physical/structural property of the polypeptide, because such specific binding relies upon the amino acid sequence of the polypeptide as well as the three-dimensional alignment of the amino acid residues. *See, e.g.*, Paul, *Fundamental Immunology*, pages 242-247 (3rd ed, 1993), attached as **Exhibit A**. As described by Paul on page 243, last full paragraph in the left column, there are two types of protein antigenic determinants (or epitopes): sequential and conformational, depending on whether the primary sequence or the three-dimensional conformation contributes the most to the

binding between the protein and its antibody. Both sequential and conformational epitopes of a protein can be predicted based on the primary amino acid sequence of a protein (*see, e.g.*, Hopp and Woods, *Pro. Natl. Acad. Sci. U.S.A.*, **78**:3824-3828, 1981, attached as **Exhibit B**; and Lewis *et al.*, *Pro. Natl. Acad. Sci. U.S.A.*, **68**:2293-2297, 1971, attached as **Exhibit C**). As such, by reciting the specific binding between the claimed polypeptides and polyclonal antibodies against a reference amino acid sequence, the pending claims set forth commonly shared structural features of the claimed genus of polypeptides.

On the other hand, commonly shared functional feature of the claimed genus of polypeptides is also provided: each is capable of conferring mitoxantrone resistance to S1-M1-80 human colon carcinoma cells when expressed in the cells. This functional feature can be readily tested by one of ordinary skill in the art using well established, routinely practiced techniques as well as according to the teaching of the present specification (*see, e.g.*, page 35 line 24 to page 37 line 2 and Example 3 on page 38).

Thus, both structural and functional features commonly shared by the claimed genus have been described in detail, which "clearly allow persons of ordinary skill in the art to recognize that [the applicant] invented what is claimed." *Vas-Cath Inc. v. Mahurkar*, 19 USPQ2d 1111, 1116 (Fed. Cir. 1991). Such description is consistent with the standards set forth in both *Lilly* and *Amgen*. Applicants therefore believe that the claimed invention within the current claim scope is properly described by the specification under 35 U.S.C. §112, first paragraph.

The withdrawal of the written description rejection is respectfully requested.

B. 35 U.S.C. §102: Anticipation

Claims 1 and 2 were rejected under 35 U.S.C. §102(a) for alleged anticipation by the Doyle *et al.* reference. Claims 1 and 2 were also rejected under 35 U.S.C. §102(e) for alleged anticipation by U.S. Patent No. 6,313,277 to Ross *et al.* Applicants respectfully traverse the rejection in light of the present amendment.

The Doyle reference was published in March 1998, whereas the Ross patent was filed on February 5, 1999, and claims priority to a provisional application filed February 5, 1998. The present application is, as the Examiner has recognized, entitled to a priority date of November 28, 1998.

As established by the Declaration of Drs. Michael Dean, Rando Allikmets, Susan Bates, and Antonio Fojo under 37 C.F.R. §1.131, including the evidence accompanying the Declaration, both filed herewith, the present inventors had in their possession well before February 5, 1998, the full length cDNA as depicted in SEQ ID NO:1 of the present application, which encodes a full length protein as depicted in SEQ ID NO:2. Despite the fact that a misreading of the DNA sequencing data led to an initial erroneous translation of a short section of the amino acid sequence, this reading error was later corrected when SEQ ID NO:1 and SEQ ID NO:2 were established for this application. Therefore, the present invention was conceived and reduced to practice by the present inventors prior to February 5, 1998.

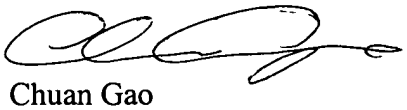
As such, Applicants submit that the Doyle *et al.* reference and the Ross patent are not available as prior art references under 35 U.S.C. §102(a) or §102(e). Accordingly, the withdrawal of the anticipation rejections is respectfully requested.

CONCLUSION

In view of the foregoing, Applicants believe all claims now pending in this Application are in condition for allowance. The issuance of a formal Notice of Allowance at an early date is respectfully requested.

If the Examiner believes a telephone conference would expedite prosecution of this application, please telephone the undersigned at 415-576-0200.

Respectfully submitted,



Chuan Gao
Reg. No. 54,111

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, Eighth Floor
San Francisco, California 94111-3834
Tel: 415-576-0200
Fax: 415-576-0300
Attachments (Exhibits A-C)
60071605 v1

FUNDAMENTAL IMMUNOLOGY

THIRD EDITION

Editor

WILLIAM E. PAUL, M.D.

Laboratory of Immunology
National Institute of Allergy and
Infectious Diseases
National Institutes of Health
Bethesda, Maryland

Raven Press New York

Raven Press, Ltd., 1185 Avenue of the Americas, New York, New York 10036

Made in the United States of America

Library of Congress Cataloging-in-Publication Data

Fundamental immunology/editor, William E. Paul.—3rd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-7817-0022-1
1. Immunology. I. Paul, William E.
[DNLM: 1. Immunity. QW 504 F9804 1993]
QR181.F84 1993
616.07'9—dc20
DNLM/DLC
for Library of Congress 93-9718
CIP

The material contained in this volume was submitted as previously unpublished material, except in the instances in which credit has been given to the source from which some of the illustrative material was derived.

Great care has been taken to maintain the accuracy of the information contained in the volume. However, neither Raven Press nor the editors can be held responsible for errors or for any consequences arising from the use of the information contained herein.

Materials appearing in this book prepared by individuals as part of their official duties as U.S. Government employees are not covered by the above-mentioned copyright.

9 8 7 6 5 4 3 2 1

groove-type sites may be derived from the same family of germline V_H genes bearing the QUPC 52 idiotype (18).

The large number of environmental carbohydrate antigens and the high degree of specificity of antibodies elicited in response to each carbohydrate antigen suggest that a tremendous diversity of antibody molecules must be available, from which some antibodies can be selected for every possible antigenic structure. In order to regulate such a diverse system, a network theory has been proposed, in which antibodies are themselves recognized as antigenic (see Chapters 12 and 24, and ref. 19), and the response to streptococcal polysaccharide is a leading example in which anti-idiotypic antibodies can be shown to regulate the response to antigen (20).

Recent studies of a series of 17 monoclonal anti-α (1 → 6) dextran hybridomas (21,22) have investigated whether the binding sites of closely related antibodies would be derived from a small number of variable region genes, for both heavy and light chains, or whether antibodies of the same specificity could derive from variable region genes with highly divergent sequences. Each monoclonal had a groove-type site that could hold six or seven sugar residues (with one exception), based on inhibition of immunoprecipitation by different length oligosaccharides. Thus, unlike monoclonals to haptenated proteins, the precise epitope could be well characterized and was generally quite similar among the entire series.

Studies of the V_κ sequences revealed that only three V_κ groups were used in these hybridomas. Use of each V_κ group correlated with the particular antigen used to immunize the animals, whether linear dextran or short oligosaccharides, so that 10 of the monoclonals from mice immunized the same way all used the same V_κ.

In contrast, the 17 V_H chains were derived from at least five different germline genes from three different V_H gene families (23). The two most frequently used germline V_H genes were found in seven and five monoclonals, respectively, with minor variations explainable by somatic mutations. Once again, V_H gene usage correlated with size of the antigen used to immunize, although the length of each CDR did not correlate with the size of the groove-type binding site. The remarkable finding is that very different V_H chains (about 50% homologous) can combine with the same V_K to produce antibody-binding sites with nearly the same size, shape, antigen specificity, and affinity. A similar phenomenon can also occur when different V_H sequences combine with different V_K sequences to produce antibodies with very similar properties. This is a result of the fact that dextran binding depends on the antigen fitting into the groove and interacting favorably with the residues forming the sides and bottom of the groove. The results indicate that divergent variable region sequences, both in and out of the complementarity-determining regions, can be folded to form similar binding site contours, which result in similar immunochemical characteristics. Similar results have been reported in other antigen–antibody systems, such as phenyloxazolone (24).

More recently, these studies were expanded to include 34 groove-type monoclonal anti-α(1 → 6) dextran-binding hybridomas (25), of which 10 used heavy chain V_H19.1.2 and eleven used V_H9.14.7. Starting with different V_H genes, these two families of monoclonals provide an experiment of nature concerning the ability of each V_H gene to combine with different light chain V_K and J_K genes, as well as heavy chain D and J_H genes to produce a groove-shaped binding site of a given specificity. In every one of these 21 monoclonals, the same light chain V_K-Ox1 gene was used, but the V_H19 family used a single J_K sequence exclusively (J_K2), while the V_H9 family included all four of the active J_K segments (J_K1, 2, 4, and 5). Similarly, the heavy chain J_H sequences of the V_H19 family were all of a single type (J_H3), while those of the V_H9 family included three types (J_H1, 2, and 3). A single D region was used by both families (DFL16.1), but the junctional sequences between V_H–D and D–J_H were different, with the V_H19 using minimal substitutions, and the V_H9 allowing more variability in junctional sequences, depending on the size of the J_H with which it was joining. Although the amino acid sequences of these two V_H genes are 73% identical, they use markedly different strategies to arrive at the same groove-type binding site with nearly identical size and specificity. The results suggest that the two heavy chain variable regions, perhaps due to their conformation, may place different structural constraints on which minigene components can successfully contribute to forming a particular site. Two different strategies for generating antibody specificity are apparent, even though the same V_K and D minigenes were used by both families. For the V_H19 family, point mutations in the CDR2 generated the α(1 → 6) dextran specificity, while the rest of the structure was held constant. For the V_H9 family, a wide variety of J_H, J_K, and V_H–D and D–J_H sequences were used to generate the groove-type site. These two blueprints for constructing a binding site may also reflect distinct cellular pathways for generating antibody diversity.

Protein and Polypeptide Antigenic Determinants

Like the proteins themselves, the antigen determinants of proteins consist of amino acid residues in a particular three-dimensional array. The residues that make contact with complementary residues in the antibody-combining site are called contact residues. To make contact, of course, these residues must be exposed on the surface of the protein, not buried in the hydrophobic core. Since the complementarity-determining residues in the hypervariable regions of antibodies have been found to span as much as 30 to 40 Å × 15 to 20 Å × 10 Å (D. R. Davies, *personal communication*), these contact

residues comprising the antigenic determinant may cover a significant area of protein surface, as now measured in a few cases by x-ray crystallography of antibody–protein antigen complexes (26–28). Looked at from another point of view, the size of the combining sites has been estimated using simple synthetic oligopeptides of increasing length, such as oligolysine. In this case, a series of elegant studies (29–31) suggested that the maximum length of chain a combining site could accommodate was six to eight residues, corresponding closely to that found earlier for oligosaccharides (14,15), discussed previously.

Several types of interactions contribute to the binding energy. Many of the amino acid residues exposed to solvent on the surface of a protein antigen will be hydrophilic. These are likely to interact with antibody contact residues via polar interactions. For instance, an anionic glutamic acid carboxyl group may bind to a complementary cationic lysine amino group on the antibody, or vice versa; or a glutamine amide side chain may form a hydrogen bond with the antibody. However, hydrophobic interactions can also play a major role. Proteins cannot exist in aqueous solution as stable monomers with too many hydrophobic residues on their surface. Those hydrophobic residues that are on the surface can contribute to binding to antibody for exactly the same reason. When a hydrophobic residue in a protein antigenic determinant or, similarly, in a carbohydrate determinant (8) interacts with a corresponding hydrophobic residue in the antibody-combining site, the water molecules previously in contact with each of them are excluded. The result is a significant stabilization of the interaction. A thorough review of these aspects of the chemistry of antigen–antibody binding has recently been published (32).

Mapping Epitopes: Conformation Versus Sequence

The other component that defines a protein antigenic determinant, besides the amino acid residues involved, is the way these residues are arrayed in three dimensions. Since the residues are on the surface of a protein, we can also think of this component as the topography of the antigenic determinant. Sela (33) divided protein antigenic determinants into two categories, sequential and conformational, depending on whether the primary sequence or the three-dimensional conformation appeared to contribute the most to binding. On the other hand, since the antibody-combining site has a preferred topography in the native antibody, it would seem *a priori* that some conformations of a particular polypeptide sequence would produce a better fit than others and therefore would be energetically favored in binding. Thus conformation or topography must always play some role in the structure of an antigenic determinant.

Moreover, when one looks at the surface of a protein in a space-filling model, one cannot ascertain the direction of the backbone or the positions of the helices (contrast Figs. 3 and 4). It is hard to recognize whether two residues that are side by side on the surface are adjacent on the polypeptide backbone or whether they come from different parts of the sequence and are brought together by the folding of the molecule. If a protein maintains its native conformation when an antibody binds, then it must similarly be hard for the antibody to discriminate between residues that are covalently connected directly and those connected only through a great deal of intervening polypeptide. Thus the probability that an antigenic determinant on a native globular protein consists of only a consecutive sequence of amino acids in the primary structure is likely to be rather small. Even if most of the determinant were a continuous sequence, other nearby residues would probably play a role as well. Only if the protein were cleaved into fragments before the antibodies were made would there be any reason to favor connected sequences.

This concept was analyzed and confirmed quantitatively by Barlow et al. (39), who examined the atoms lying within spheres of different radii from a given surface atom on a protein. As the radius increases, the probability that all the atoms within the sphere will be from the same continuous segment of protein sequence decreases rapidly. Correspondingly, the fraction of surface atoms that would be located at the center of a sphere containing only residues from the same continuous segment falls dramatically as the radius of the sphere increases. For instance, for lysozyme, with a radius of 8 Å, fewer than 10% of the surface residues would lie in such a "continuous patch" of surface. These are primarily in regions that protrude from the surface. With a radius of 10 Å, almost none of the surface residues fall in the center of a continuous patch. Thus for a contact area of about 20 Å × 25 Å, as found for a lysozyme–antibody complex studied by x-ray crystallography, none of the antigenic sites could be completely continuous segmental sites.

Antigenic sites consisting of amino acid residues that are widely separated in the primary protein sequence but brought together on the surface of the protein by the way it folds in its native conformation have been called "assembled topographic" sites (40,41) because they are assembled from different parts of the sequence and exist only in the surface topography of the native molecule. By contrast, the sites that consist of only a single continuous segment of protein sequence have been called "segmental" antigenic sites (40,41).

In contrast to T cell recognition of "processed" fragments retaining only primary and secondary structures, the evidence is overwhelming that most antibodies are made against the native conformation when the native protein is used as immunogen. For instance, antibodies to native staphylococcal nuclease were found to have about a 5000-fold higher affinity for the native protein

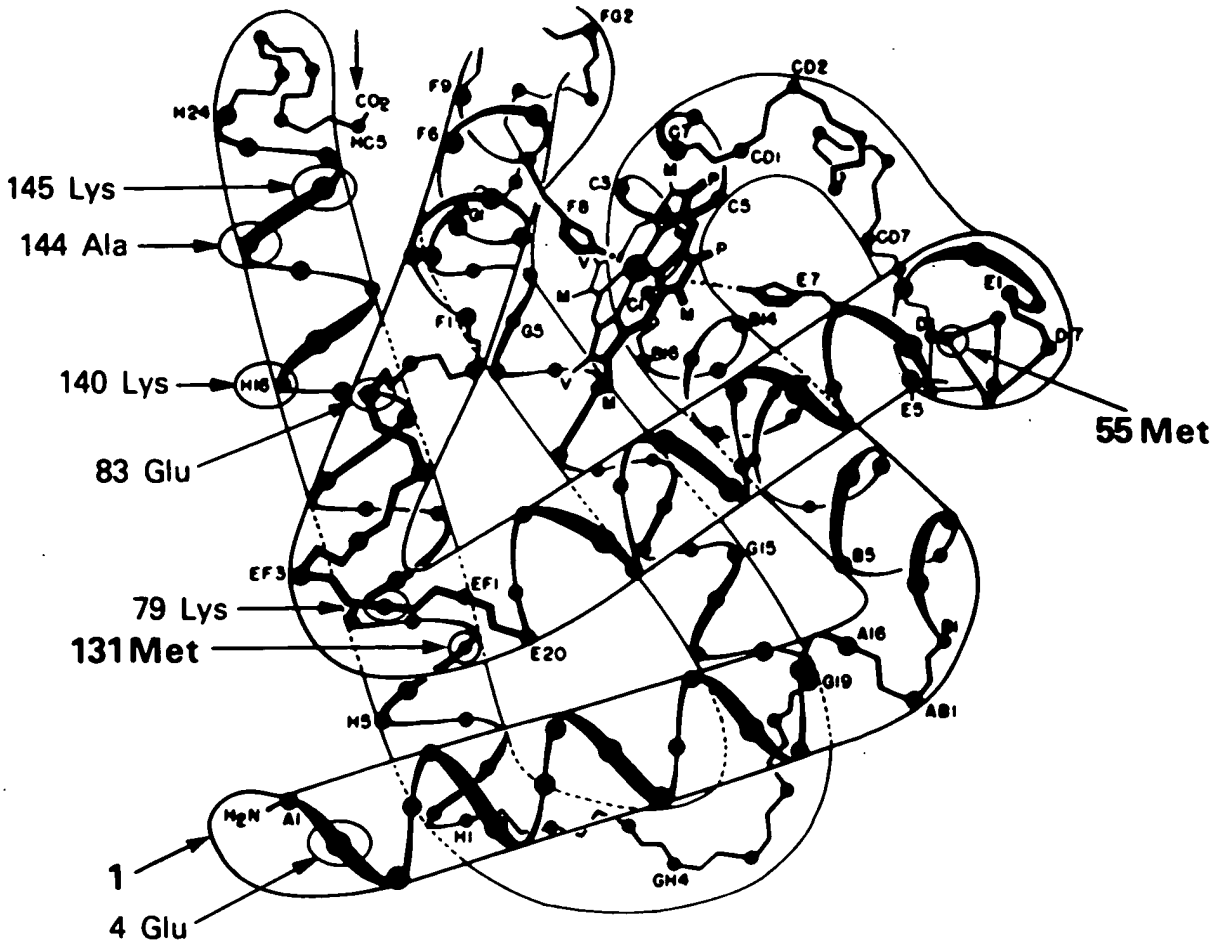


FIG. 3. Artist's representation of the polypeptide backbone of sperm whale myoglobin in its native three-dimensional conformation. The α helices are labeled A through H from the amino terminal to the carboxyl terminal. Side chains are omitted, except for the two histidine rings (F8 and E7) involved with the heme iron. Methionines at positions 55 and 131 are the sites of cleavage by cyanogen bromide (CNBr), allowing myoglobin to be cleaved into three fragments. Most of the helicity and other features of the native conformation are lost when the molecule is cleaved. A less drastic change in conformation is produced by removal of the heme to form apomyoglobin, since the heme interacts with several helices and stabilizes their positions relative to one another. The other labeled residues (4 Glu, 79 Lys, 83 Glu, 140 Lys, 144 Ala, and 145 Lys) are residues that have been found to be involved in antigenic determinants recognized by monoclonal antibodies (34). Note that cleavage by CNBr separates Lys 79 from Gly^{34} 4 and separates Glu 83 from Ala 144 and Lys 145. The "sequential" determinant of Koketsu and Atassi (35) (residues 15 to 22) is located at the elbow, **lower right,** from the end of the A helix to the beginning of the B helix. (Adapted from ref. 36.)

than for the corresponding polypeptide on which they were isolated (by binding to the peptide attached to Sepharose) (42). An even more dramatic example is that demonstrated by Crumpton (43) for antibodies to native myoglobin or to apomyoglobin. Antibodies to native ferric myoglobin produced a brown precipitate with myoglobin, an indication that the heme was still in the protein in what was, at least approximately, its native environment. Such antibodies did not bind well to the apomyoglobin, which, without the heme, has a slightly altered conformation. On the other hand, antibodies to the apomyoglobin, when mixed with native (brown) myoglobin, produced a white precipitate. These antibodies so strongly favored the conformation of apomyoglobin, from which the heme was excluded, that they trapped those molecules that vibrated toward that conformation and pulled the equilibrium state over to the apo form. One could almost say, figuratively, that the antibodies squeezed the heme out of the myoglobin. Looked at thermodynamically, it is clear that the conformational preference of the antibody for the apo versus native forms, in terms of free energy, had to be greater than the free energy of binding of the heme to myoglobin. Thus, in general, antibodies are made that are very specific for the conformation of the protein used as immunogen.

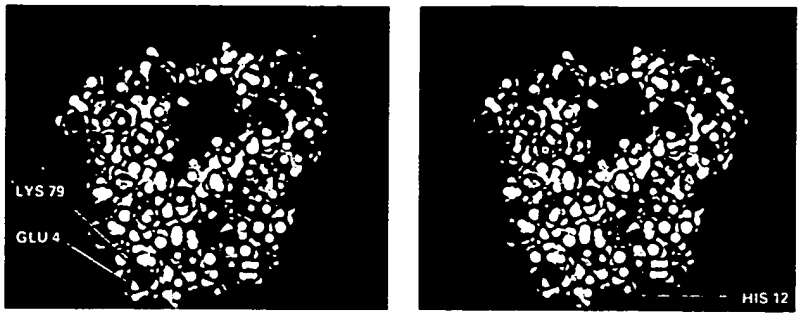


FIG. 4. Stereoscopic views of a computer-generated space-filling molecular model of sperm whale myoglobin, based on the Takano (37) x-ray diffraction coordinates. This orientation, which corresponds to that in Fig. 3, is arbitrarily designated the "front view." The computer method was described by Feldmann et al. (38). The heme and aromatic carbons are *shaded darkest*, followed by carboxyl oxygens, then other oxygens, then primary amino groups, then other nitrogens, and finally side chains of aliphatic residues. The backbone and the side chains of nonaliphatic residues, except for the functional groups, are shown in *white*. Note that the direction of the helices is not apparent on the surface, in contrast to the backbone drawing in Fig. 3. The residues Glu 4, Lys 79, and His 12 are believed to be part of a topographic antigenic determinant recognized by a monoclonal antibody to myoglobin (34). This stereo pair can be viewed in three dimensions using an inexpensive stereoviewer such as the "stereoscopes" sold by Abrams Instrument Corp., Lansing, MI, or Hubbard Scientific Co., Northbrook, IL. (Adapted from ref. 34.)

A number of methods have been used to identify the antigenic determinants bound by particular antibodies made against a protein. Binding to cleavage fragments and short synthetic peptides from the protein sequence has been the most widely used approach. The synthetic peptides may be made by conventional solid-phase peptide synthesis (44) or by methods designed to make large numbers of peptides for screening. In one such method, multiple peptides are made simultaneously in separate polypropylene mesh "tea-bags" that can be put through the common steps in the sequence together and separated only for the different amino acid coupling steps (45). In another method, the peptides are synthesized on the tips of plastic pins inserted in the wells of 96-well microtiter plates in such a way that these can then be used for solid-phase binding assays of antibodies without ever cleaving the peptide off the plastic support (46). These two methods especially lend themselves to studying multiple variants of the natural sequence to identify the residues critical for antibody binding. Usually, the longer the peptides, the more that specificity can be confidently determined, as short peptides of only six to eight amino acid residues often manifest nonspecific binding (47). If the synthetic peptides correspond to segments of the protein antigen sequence, as is most common, then the use of peptides is limited to identifying the structures bound by antibodies specific for segmental antigenic sites.

To identify assembled topographic sites, more complex approaches have been necessary. The earliest was the use of natural variants of the protein antigen with known amino acid substitutions, where such evolutionary variants exist (40). Thus substitution of different amino acids in proteins in the native conformation can be examined. The use of this method, which is illustrated later, is limited to studying the function of amino acids that vary among homologous proteins, that is, those that are polymorphic. It may now be extended to other residues by use of site-directed mutagenesis. A second method is to use the antibody that binds to the native protein to protect the antigenic site from modification (48) or proteolytic degradation (49). A related but less sensitive approach makes use of competition with other antibodies (50–52). A third approach, taking advantage of the capability of producing thousands of peptides on a solid-phase surface for direct binding assays (46), is to study binding of a monoclonal antibody to every possible combination of six amino acids (46). If the assembled topographic site can be mimicked by a combination of six amino acids not corresponding to any continuous segment of the protein sequence but structurally resembling a part of the surface, then one can produce a "mimotope" defining the specificity of that antibody (46).

Myoglobin also serves as a good model protein antigen for studying the range of variation of antigenic determinants from those that are more sequential in nature to those that do not even exist without the native conformation of the protein (Fig. 3). A good example of the first, more segmental type of determinant is that consisting of residues 15 to 22 in the amino terminal portion of the molecule. Crumpton and Wilkinson (53) first discovered that the chymotryptic cleavage fragment consisting of residues 15 to 29 had antigenic activity for antibodies raised to either native or apomyoglobin. Synthetic peptides corresponding to the shorter sequence 15 to 22 were then found by two groups (35,54) to bind antibod-

ies made to native sperm whale myoglobin, even though the synthetic peptides were only 7 to 8 residues long. Peptides of this length do not spend much time (in solution) in a conformation corresponding to that of the native protein. On the other hand, these synthetic peptides had a several hundred-fold lower affinity for the antibodies than did the native protein. Thus, even if most of the determinant was included in the consecutive sequence 15 to 22, the antibodies were still much more specific for the native conformation of this sequence than for the random conformation peptide. Moreover, there was no evidence to exclude the participation of other residues, nearby on the surface of myoglobin but not in this sequence, in the antigenic determinant.[1]

A good example of the importance of secondary structure is the case of the loop peptide (residues 64 to 80) of hen egg-white lysozyme (59). This loop in the protein sequence is created by the disulfide linkage between cysteine residues 64 and 80 and has been shown to be a major antigenic determinant for antibodies to lysozyme (59). The isolated peptide 60 to 83, containing the loop, binds antibodies with high affinity, but opening of the loop by cleavage of the disulfide bond destroys most of the antigenic activity for antilysozyme antibodies (59).

At the other end of the range of conformational requirements are those determinants involving residues far apart in the primary sequences that are brought close together on the surface of the native molecule by its folding in three dimensions. Myoglobin also provides a good example of these determinants, which are called assembled topographic determinants (40,41). Of six monoclonal antibodies to sperm whale myoglobin studied by Berzofsky et al. (34,60), none bound to any of the three cyanogen bromide cleavage fragments of myoglobin that together span the whole sequence of the molecule. Therefore these monoclonal antibodies (all with affinities between 2×10^8 and 2×10^9 M^{-1}) were all highly specific for the native conformation. These were studied by comparing the relative affinities for a series of native myoglobins from different species with the known amino acid sequences of these myoglobins. With the myoglobins available, this approach allowed the definition of some of the residues involved in binding to three of these antibodies. The striking result was that two of these three monoclonal antibodies were found to recognize topographic determinants, as defined previously. One recognized a determinant including Glu 4 and Lys 79, which are on the A helix and E–F corner of the myoglobin molecule but come within about 2 Å of each other to form a salt bridge in the native molecule (Fig. 4). The other antibody recognized a determinant involving Glu 83 in the E–F corner, and Ala 144 and Lys 145 on the H helix of the myoglobin molecule (Fig. 5). Again, these are far apart in the primary sequence but are brought within 12 Å of each other by the folding of the molecule in its native conformation. Similar examples have recently been reported for monoclonal antibodies to human myoglobin (61) and to lysozyme (50). Other examples of such conformation-dependent antigenic determinants have been suggested using conventional antisera to such proteins as insulin (62), hemoglobin (63), tobacco mosaic virus (64), and cytochrome *c* (65). Moreover, the crystallographic structures of lysozyme–antibody (26,28) and neuraminidase–antibody (27) complexes show clearly that, in both cases, the epitope bound is an assembled topographic site.

How frequent are antibodies specific for topographic determinants compared to those that bind consecutive sequences when conventional antisera are examined? This question was studied by Lando et al. (66), who passed goat, sheep, and rabbit antisera to sperm whale myoglobin over columns of Sepharose-coupled cyanogen bromide cleavage fragments of myoglobin, together spanning the whole sequence. The antisera were passed sequentially, and repeatedly, over each of the three columns until no more antibodies could be removed. Nevertheless, 30% to 40% of the antibodies originally present in each serum remained after this treatment. These antibodies still bound to the native myoglobin molecule with high affinity but did not bind to any of the fragments in solution by radioimmunoassay. Thus, in four of four antimyoglobin sera tested, 60% to 70% of the antibodies could bind peptides and 30% to 40% could bind only native-conformation intact protein.

On the basis of studies such as these, it has been suggested that much of the surface of a protein molecule may be antigenic (40,67) but that the surface can be divided up into antigenic domains (34,57,58,61). Each of these domains consists of many overlapping determinants recognized by different antibodies.

An additional interesting point can be made from the above studies about the topography of protein antigenic determinants. If one examines the topographic determinant consisting of sperm whale myoglobin residues 83, 144, and 145, shown in stereo in Fig. 5, it is apparent that they are on both lips of a deep crevice or concavity in the protein surface. It is possible, although not yet demonstrated, that a complementary protuberance in the antibody-combining site actually inserts into this cavity. From the studies of myeloma proteins that bind small haptens or carbohydrates, we are accustomed to think of the antigen being engulfed by a cavity or crevice

[1] This is the only segmental antigenic determinant of myoglobin that has clearly been confirmed by more than one independent group of investigators. Crumpton and Wilkinson (53) did measure antigenic activity for a chymotryptic fragment 147 to 153 that overlaps one of the other reported sequential determinants (55). However, two of the other reported sequential determinants (55), corresponding to residues 56 to 62 and 94 to 100, have not been reproducible when tested with other antisera, even raised in the same species (56). For related studies, see refs. 57 and 58.

FIG. 5. Stereoscopic view of computer-generated space-filling model of the left view of sperm whale myoglobin, turned 90° relative to the view in Fig. 4. Methods and shading are as indicated in Fig. 4. The residues Glu 83, Ala 144, and Lys 145 are believed to be part of a topographic antigenic determinant recognized by a second monoclonal antibody to myoglobin (34). Note the concavity in the surface of the molecule in the middle of this determinant, between Glu 83 and Ala 144/Lys 145. (Adapted from ref. 34.)

on the antibody (68). However, for globular protein antigens binding to globular protein antibodies, the situation is more structurally symmetrical (and antigen–antibody binding is also thermodynamically symmetrical). Thus it is just as possible for a convexity on the antibody to insert into a concavity on the antigen as it is for the more conventional model to occur of a convexity on the antigen inserting into a concavity on the antibody. Now that monoclonal antibodies specific for protein antigens are available, we may encounter both types of cases. The determinant depicted in Fig. 5 might be such a case. In the three published crystal structures of protein antigen–antibody complexes, the contact surfaces were broad, with local complementary pairs of concave and convex regions in both directions (26–28). However, when we limit ourselves to antigenic sites defined with short peptides, which tend to identify sites that protrude from the surface of the antigen (39,69), we are likely to see a bias toward situations in which the antigen is convex and the antibody surface concave.

Further information on the subjects discussed in this section is available in the reviews by Sela (33), Crumpton (43), Reichlin (70), Kabat (68), Benjamin et al. (40), Berzofsky (41), and Getzoff et al. (32).

Conformational Equilibria of Protein and Peptide Antigenic Determinants

We have already referred to the fact that antibodies to a native protein have higher affinity for the native conformation than for other conformations of fragments or denatured molecules. Similarly, antibodies raised against fragments or denatured molecules generally have higher affinities for these forms than for the native conformation. In this section we discuss possible mechanisms for these affinity differences and explore how these can be used to advantage to study the conformational equilibria of proteins and peptides.

There are several possible mechanisms to explain why an antibody specific for a native protein will bind a peptide fragment in random conformation with lower affinity. Of course, the peptide may not contain all the contact residues of the antigenic determinant, so that the binding energy would be lower. However, for cases in which all the residues in the determinant are present in the peptide, several mechanisms still remain. First, the affinity may be lower because the topography of the residues in the peptide may not produce as complementary a fit in the antibody-combining site as the native conformation would. Second, it is possible that the apparent affinity is reduced because only a small fraction of the peptide molecules are in a nativelike conformation at any time. This model assumes that the antibody binds only those peptide molecules that are in the native conformation. Since the concentration of these is lower than the total peptide concentration by a factor that corresponds to the conformational equilibrium constant of the peptide, the apparent affinity is also lower by this factor. This model is analogous to an allosteric model. A third, intermediate hypothesis would suggest that initial binding of the peptide in a nonnative conformation occurs with submaximal complementarity and is followed by an intramolecular conformational change in the peptide to achieve energy minimization by assuming a nativelike conformation. This third hypothesis corresponds to an induced fit model. The loss of affinity is due to the energy required to change the conformation of the peptide, which in turn corresponds to the conformational equilibrium constant in the second hypothesis. To some extent these models could be distinguished kinetically, since the first hypothesis predicts a faster "on" rate and a faster "off" rate than does the second hypothesis (71).

Although not the only way to explain the data, the second hypothesis is useful because it provides a method to estimate the conformational equilibria of proteins and peptides (42,72). The method assumes the second hypothesis, which can be expressed as follows:

Proc. Natl. Acad. Sci. USA
Vol. 78, No. 6, pp. 3824–3828, June 1981
Immunology

Prediction of protein antigenic determinants from amino acid sequences

(hydrophilicity analysis/protein conformation)

THOMAS P. HOPP AND KENNETH R. WOODS

The Lindsley F. Kimball Research Institute, The New York Blood Center, 310 East 67th Street, New York, New York 10021

Communicated by Bruce Merrifield, March 2, 1981

ABSTRACT A method is presented for locating protein antigenic determinants by analyzing amino acid sequences in order to find the point of greatest local hydrophilicity. This is accomplished by assigning each amino acid a numerical value (hydrophilicity value) and then repetitively averaging these values along the peptide chain. The point of highest local average hydrophilicity is invariably located in, or immediately adjacent to, an antigenic determinant. It was found that the prediction success rate depended on averaging group length, with hexapeptide averages yielding optimal results. The method was developed using 12 proteins for which extensive immunochemical analysis has been carried out and subsequently was used to predict antigenic determinants for the following proteins: hepatitis B surface antigen, influenza hemagglutinins, fowl plague virus hemagglutinin, human histocompatibility antigen HLA-B7, human interferons, *Escherichia coli* and cholera enterotoxins, ragweed allergens Ra3 and Ra5, and streptococcal M protein. The hepatitis B surface antigen sequence was synthesized by chemical means and was shown to have antigenic activity by radioimmunoassay.

The elucidation of protein antigenic structures is presently a difficult, uncertain, and time-consuming task. To precisely delineate antigenic determinants, it is necessary to prepare a large number of well-characterized chemical derivatives and peptide fragments from the original protein antigen and then to test these derivatives for immunological activity (1, 2). Alternatively, a homologous series of proteins may be used to assess the influence of particular amino acid substitutions, thereby implicating certain regions as antigenic determinants (3, 4); this approach requires knowledge of complete primary structures for a number of proteins before the immunological results can be interpreted. Despite the laboriousness of available approaches, the complete antigenic structures have been elucidated for a small number of proteins, and partial information is available for many others.

As more information becomes available on protein antigens, it should be possible to use this information to predict the locations of antigenic determinants before any immunological testing has been carried out. In recent years a number of systems have been developed to predict protein conformational features from amino acid sequences (5–8), but none of these were specifically oriented to the prediction of antigenic determinants. Therefore, we sought a method that was not predicated upon predictions of particular structural features but rather sought a simple correlation with surface location of stretches of peptide chain and the likelihood of antibody binding. A guiding principle was the notion that many surface oriented regions are *nonantigenic* (1). This led us to take an empirical approach in our analysis and to arbitrarily manipulate the emphasis placed on certain amino acids in order to find a particular kind of sequence that is favored for antibody binding (which may not strictly depend on the hydrophilicity of the sequence). The present report describes a system that uses a simplified method to successfully predict antigenic determinants, given the amino acid sequence of a protein and no other information.

METHOD

Previous investigations have demonstrated that antigenic determinants are surface features of proteins and indicate that they are frequently found on regions of a molecule that have an unusually high degree of exposure to solvent—i.e., regions which project into the medium (for reviews, see refs. 1 and 3). This, together with the fact that charged, hydrophilic amino acid side chains are common features of antigenic determinants, led us to investigate the possibility that at least some antigenic determinants might be associated with stretches of amino acid sequence that contain a large number of charged and polar residues and are lacking in large hydrophobic residues. A suitable means of methodically searching for such regions was found by combining a method like that of Chou and Fasman (5), in which numerical values for amino acids are repetitively averaged over the length of a polypeptide chain, with a set of values expressing the relative hydrophilicity of each amino acid. Suitable values were available in the solvent parameters assigned by Levitt (6), which are derivatives of the hydrophobicity values of Nozaki and Tanford (9).

In Table 1 are listed the numerical values (hydrophilicity values) assigned to the 20 amino acids commonly found in proteins. In the first column, the values of Levitt (6) are listed, whereas the second column lists the values that were finally chosen for our hydrophilicity calculations. The values were generally retained as expressed by Levitt; however, changes in the values for proline, asparatic acid, and glutamic acid improve the prediction results, as explained later. Hydrophilicity analysis of a protein is carried out by the following method.

Each amino acid in the sequence of the protein is assigned its hydrophilicity value, then these values are repetitively averaged down the length of the polypeptide chain, generating a series of local hydrophilicity values. The number of hydrophilicity values that are averaged at each repetition is arbitrary, and we chose groups of six for our initial studies because this is the approximate size of an antigenic determinant (1, 10). Once the complete set of averaged values is obtained, the list is scanned to locate the highest value. According to the studies presented here, this high point will invariably lie within or be immediately adjacent to one of that protein's antigenic determinants.

A useful way of recording the results of this analysis is to produce a plot of hydrophilicity value versus sequence position.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: αAbu, α-aminobutyric acid; HBsAg, hepatitis B surface antigen.

Table 1. Hydrophilicity values

Amino acid	s,* kcal/mol	Hydrophilicity value
Arginine	3.0	3.0
Aspartic acid	2.5	3.0
Glutamic acid	2.5	3.0
Lysine	3.0	3.0
Serine	0.3	0.3
Asparagine	0.2	0.2
Glutamine	0.2	0.2
Glycine	0.0	0.0
Proline	−1.4	0.0
Threonine	−0.4	−0.4
Alanine	−0.5	−0.5
Histidine	−0.5	−0.5
Cysteine	−1.0	−1.0
Methionine	−1.3	−1.3
Valine	−1.5	−1.5
Isoleucine	−1.8	−1.8
Leucine	−1.8	−1.8
Tyrosine	−2.3	−2.3
Phenylalanine	−2.5	−2.5
Tryptophan	−3.4	−3.4

* Solvent parameter values assigned by Levitt (6).

Fig. 1, the hexapeptide analysis of sperm whale myoglobin, is illustrative. The high point of the profile, at position 60.5, falls within myoglobin antigenic site 2 (1). Several findings which proved to be generally true with other proteins can be seen in the myoglobin plot. First, not all antigenic determinants are associated with high points of hydrophilicity (for example, antigenic site 4, residues 113 through 119); second, not all high points are associated with antigenic determinants (position 79.5). The one correlation which has been upheld in myoglobin and the other proteins that we tested, is that one antigenic determinant is consistently located at the point of maximum hydrophilicity.

Computerization. To facilitate the analysis of large quantities of sequence information, our procedure was encoded in a FORTRAN program and run in a PDP 11/70 computer, and the resulting data was plotted with a Tektronix automatic plotting device.

List of Antigenic Determinants. Proteins with known antigenic determinants were considered to belong to one of two groups. Group 1, proteins whose antigenic structures are nearly or completely solved includes: (*i*) sperm whale myoglobin, with antigenic determinants at residues 15–22 (site 1), 56–62 (site 2), 94–99 (site 3), 113–119 (site 4), and 145–151 (site 5) (1); (*ii*) chicken lysozyme, with antigenic determinants including residues 5, 7, 13, 14, 33, 34, 62, 87, 89, 93, 96, 97, 113, 114, 116, and 125 (2); (*iii*) the ferredoxin from *Clostridium pasteurianum*, with antigenic determinants encompassing residues 1–7 and 51–55 (11); (*iv*) horse heart cytochrome *c*, with antigenic residues at positions 47, 58–62, 88–92, and 96 (4, 12); and (*v*) bovine myelin basic protein, with determinants in regions 64–73, 74–85, 113–121, and 153–166 (13, 14). Group 2, proteins for which partial information is available, comprises: (*i*) human hemoglobin β chains, with antigenic residues at 6, 16–23, 52, 68, 73, and 102 (3, 15, 16); (*ii*) the tobacco mosaic virus (vulgare) coat protein, with antigenic determinants at positions 62–68, 108–113, and 153–158 (17–19); (*iii*) human IgG heavy chain constant regions (each of the three constant domains of the Eu myeloma protein was considered as an individual protein), with antigenic determinants localized to position 214 of the CH1 domain, positions 296 and 309 of the CH2 domain, and 355 to 358 of the CH3 domain (20); (*iv*) bovine α-lactalbumin, where antigenic determinants have been located within residues 10–18, 60–80, 91–94, and 105–117 (unpublished data); and (*v*) leghemoglobin a from the soybean, with antigenic sites within residues 15–23, 52–59, 92–98, 107–116, and 132–142 (21).

Evaluating Predictions. An antigenic determinant was considered to be correctly identified by a prediction point if that point fell within the determinant, directly on a single antigenic residue, or within two residues (inclusive) on either side of any antigenic residue. This inclusion of a two residue "buffer zone" around antigenic sites is acceptable because much of the available information implicates single residues as antigenic sites, although in most cases these residues probably comprise part

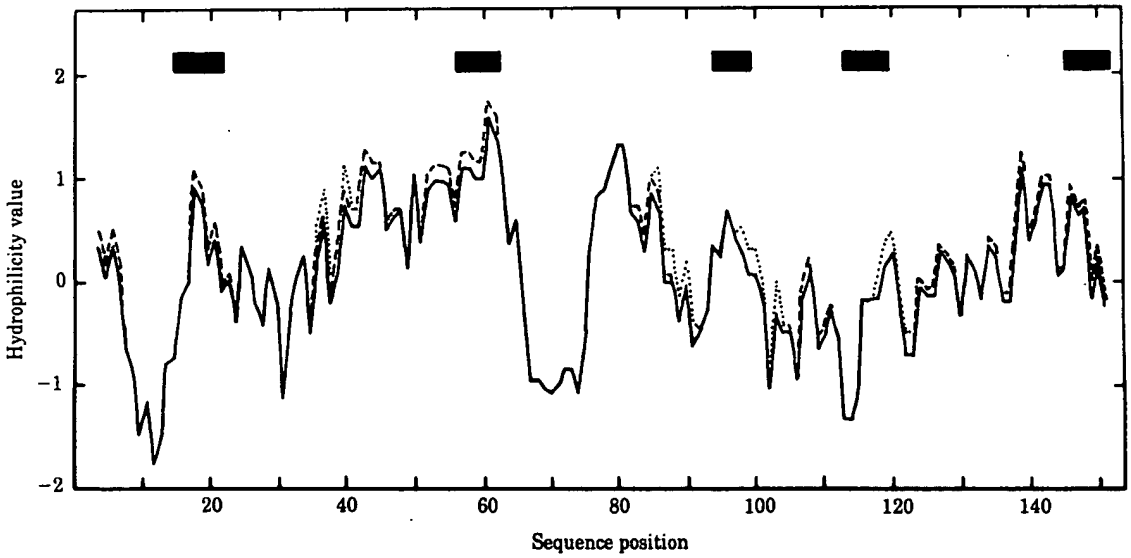


FIG. 1. Hexapeptide profile of sperm whale myoglobin. The averaged antigenicity values are plotted versus position along the amino acid sequence. The *x* axis contains 153 increments, each representing an amino acid in the sequence of myoglobin. The *y* axis represents the range of hydrophilicity values (from 3 to −3.4). The data points are plotted at the center of the averaging group from which they were derived. ■, Known antigenic determinants of myoglobin; ——, profile obtained by assigning the "solvent parameter" values of Levitt (6) to each amino acid; ----, profile obtained when the values for aspartic acid and glutamic acid were raised to 3.0; ·····, profile obtained when proline was assigned the value of 0.

Table 2. Prediction success with hexapeptide maximum point

	Correct	Wrong	Unknown
Original values	7	1	4
Asp, Glu = 3	8	0	4
Asp, Glu = 3, Pro = 0	10	0	2

of a larger site that includes several residues immediately adjacent to them in the sequence. Furthermore, any experiments designed to test the validity of these antigenic determinant predictions would be expected to include a number of residues on either side of the predicted point, in which case an overlap with the antigenic determinant would always be guaranteed.

Owing to the limited information available on the antigenic structures of some of the proteins used in this study, it was not always possible to definitely assess the correctness of a given prediction. Therefore, the proteins of groups 1 and 2 were treated differently in generating the information shown in Tables 2 and 3.

For group 1 proteins, a prediction was considered correct (Tables 2 and 3, column 1) if it successfully located an antigenic determinant or wrong (column 2) if it missed. With group 2 proteins, however, it is possible that a predicted point that misses *known* antigenic determinants may be indicating an antigenic determinant that is currently undiscovered. Therefore, for these proteins, predictions were considered to be correct (column 1) if they hit a known determinant or unknown if they missed (because they may yet prove to be hits).

Adjustment of Aspartic Acid, Glutamic Acid, and Proline Values. Table 2 shows the effect of increasing the values for these three amino acids from the original values given by Levitt. Increasing aspartic acid and glutamic acid from 2.5 to 3.0 eliminated the one wrong prediction and caused an elevation of the plots in many regions where antigenic determinants are known to exist (e.g., myoglobin sites 1, 2, and 5). There was no change in the number of unknown predicted points in the group 2 proteins (column 3), although the new values tended to elevate the profiles in the locations of the known determinants in these proteins. Next, the value for proline was raised to zero, and the hexapeptide analyses were repeated. The result is shown in line 3 of Table 2: two of the proteins that had given unknown predicted points now resulted in correct predictions.

The two remaining proteins with unknown prediction points are unusual, and it may not be worthwhile to attempt to bring them into the "correct" group by making further changes in amino acid values. For one of the two, the CH2 region of IgG, only 2 out of 109 residues are presently known to be antigenically active, and it may be possible that the predicted point may be indicating an undiscovered antigenic determinant. In the case of leghemoglobin a, the investigators indicate that they have not tested the antigenic activity of the predicted region (21).

Table 3. Effect of averaging group length on predictions by the maximum point

	Correct	Wrong	Unknown	C/C+W, %*
Dipeptide	23	17	18	58
Tripeptide	10	5	3	67
Tetrapeptide	9	3	4	75
Pentapeptide	8	2	2	80
Hexapeptide	10	0	2	100
Heptapeptide	7	3	2	70
Octapeptide	6	2	4	75
Nonapeptide	5	3	4	63
Decapeptide	5	2	5	71

* Percentage of correct assignments when considering only proteins of group 1. C, correct; W, wrong.

The Effect of Averaging Group Length. When only two amino acids are averaged at a time, the data plot is erratic, and the great variation in hydrophilicity over short lengths of peptide tends to obscure the general trend of the values. In addition, dipeptide analysis results in multiple identical high points because any pair of charged residues will yield the maximum value of 3.0. The results of this can be seen in line one of Table 3. For the 12 proteins analyzed, a total of 58 identical high points were obtained, and only 23 of these were associated with known antigenic determinants. Moreover, dipeptide analysis resulted in 17 wrong predictions.

Multiple identical high points continues to be a problem for tri- and tetrapeptide analysis; it finally disappears at the pentapeptide level (and higher), yielding a single predicted point for each of the 12 proteins. Although it is attractive to consider a method like the di-, tri-, or tetrapeptide analysis, which can predict more than one determinant per molecule, it seems more important to eliminate as many wrong predictions as possible because they reduce confidence in any given predicted antigenic determinant. As averaging group length increases, the number of wrong predictions decreases to a minimum of zero for hexapeptide analysis (Table 3). Comparison of data plots for various averaging group lengths suggested a reason for this. In going from di- to tetra- to hexapeptide analysis, the plots became less chaotic and the local hydrophilicity trend became more apparent. In going from hexa- to octa- to decapeptide analysis, the plots became even smoother. However, wrong predictions appeared again, and there was an increase in unknown predictions, whereas correct predictions fell from 10 to a low of 5 for nona- and decapeptide analysis. The reason for this may be that the regions of high hydrophilicity that are recognized well by the hexapeptide analysis begin to be obscured when longer averaging groups were used, due to their being combined with adjacent regions of low hydrophilicity.

Second and Third Highest Points. In order to assess the generality of the predictive value of high points, the success of the second and third highest points was considered. These points were only selected from the subset of points that had at least three amino acid positions between them and the highest (or second highest) point. This resulted in the second and third highest points always occurring in their own individual peak of hydrophilicity and the elimination of redundant prediction of antigenic determinants. However, neither the second nor the third highest points gave highly reliable prediction results. Although the correlation of predicted points with antigenic determinants seems to be significant in both cases (25% for the second and 33% for the third), the number of wrong predictions (33% in each case) severely limits the usefulness of these points for prediction of antigenic determinants of unknown proteins. These points are probably worthy of consideration in cases where immunochemical testing is used to verify the predictions because (by ignoring unknown predictions) they represent a 43% and 50% chance of a correct prediction, respectively.

Predictions for Uncharacterized Protein Antigens. We have applied our procedure to a number of proteins for which the location of an antigenic determinant may be of particular interest (Table 4). Several of the sequences listed in Table 4 are longer than six amino acids. In those cases, there are two or more adjacent sets of amino acids that result in identical average hydrophilicity values. Synthesis of short peptides should verify that these sequences are in, or immediately adjacent to, antigenic determinants.

To this end, we have recently used the Merrifield procedure to synthesize a peptide having the sequence αAbu-αAbu-Thr-

Table 4. Protein sequences with greatest average hydrophilicity*

Protein	Sequence
HBsAg (22)	141-Lys-Pro-Thr-Asp-Gly-Asn
Influenza hemagglutinins	
A/Victoria/3/75 strain (23)	171-Asn-Asp-Asn-Ser-Asp-Lys
A/Aichi/2/68 strain (24)	88-Val-Glu-Arg-Ser-Lys-Ala
Fowl plague virus hemagglutinin (25)	97-Glu-Arg-Arg-Glu-Gly-Asn
Human histocompatibility antigen HLA-B7 (26)	43-Pro-Arg-Glu-Glu-Pro-Arg
Human interferons	
Fibroblast (27)	103-Glu-Glu-Lys-Leu-Glu-Lys-Glu-Asp
Leukocyte I (28)	160-Glu-Arg-Leu-Arg-Arg-Lys-Glu
Leukocyte A (29)	131-Lys-Glu-Lys-Lys-Tyr-Ser
E. coli enterotoxins	
Heat labile (30)	66-Glu-Arg-Met-Lys-Asp-Thr
Heat stable (31)(two identical peaks)	26-Asp-Ser-Ser-Lys-Glu-Lys 46-Ser-Glu-Lys-Lys-Ser-Glu
Cholera toxin β chain (32)	79-Glu-Ala-Lys-Val-Glu-Lys
Streptococcal M protein (33)	58-Arg-Lys-Ala-Asp-Leu-Glu-Lys
Ragweed allergens	
Ra3 (34)	88-Cys-Thr-Lys-Asp-Gln-Lys
Ra5 (35)	40-Ser-Lys-Lys-Cys-Gly-Lys
Semliki Forest virus membrane proteins (36)	
E1	70-Thr-Lys-Glu-Lys-Pro-Asp
E2	246-Asp-Glu-Pro-Ala-Arg-Lys
E3	40-Glu-Asp-Asn-Val-Asp-Arg

* For each protein listed, the sequence of amino acids having the greatest average hydrophilicity value is shown; the number before the sequence indicates the position of the first amino acid in the group.

Lys-Pro-Thr-Asp-Gly-Asn-αAbu-Thr-αAbu (αAbu = α-amino butyric acid, replacing Cys) corresponding to residues 138–149 of the hepatitis B surface antigen (HBsAg) protein, and tested it for antigenic activity. The peptide side chains were deprotected under conditions where the peptide remained attached to the polystyrene beads (21). The peptidyl beads were then used to replace the polystyrene beads normally used in the Ausria II radioimmunoassay for HBsAg (Abbott), yielding a clearly positive binding affinity for ^{125}I-labeled anti-HbsAg antibodies. Beads without peptide, or peptidyl beads in which the side chain protecting groups had not been removed, did not bind significant ^{125}I-labeled anti-HBsAg antibody. Details of these experiments will be published elsewhere.

DISCUSSION

The studies described demonstrate the usefulness and limitations of antigenic determinant prediction by hydrophilicity analysis. The peak hexapeptide prediction value is highly successful, yielding no wrong assignments in 12 proteins; only lack of information on 2 of the 12 proteins makes it unclear whether the present method has a 100% success rate. On the other hand, the second and third highest peaks result in a mixture of correct and incorrect assignments and therefore, are less useful as predictors of antigenic determinants. It is clear by inspection of the data plots that some antigenic determinants are not correlated with hydrophilicity, although there does seem to be a correlation of many antigenic determinants with local upspikes of the hydrophilicity profile. This suggests that our present method may be a good basis on which to superimpose other types of information that boost the values of these low peaks. For example, it may be possible to improve prediction success by considering currently available methods for predicting secondary structure, particularly β bends.

Our method bears some resemblance to the procedure reported by Rose and Roy for predicting protein packing by hydrophobicity analysis (8), but it also has distinct differences that make it a better system for locating antigenic determinants. Because their approach utilizes the hydrophobicity values of Nozaki and Tanford (9) without the adjustments introduced by Levitt (6), the values for all hydrophilic amino acids are identical (i.e., 0), whereas the corresponding values used in our procedure range from 0.2 to 3.0. This results in a strong influence by the charged amino acids and an intermediate effect for neutral polar amino acids. Furthermore, Rose and Roy use a least-squares fitting of data to a quadratic polynomial with a seven-point moving window rather than hexapeptide averaging. This results in greater smoothing of the profile and end effects. Both of these qualities seem to decrease the potential usefulness for antigenic determinant prediction. In contrast, our method depends upon simpler calculations and a shorter averaging-group length and is capable of considering all amino acids from the amino-terminal to the carboxyl-terminal residue.

Finally, it should be emphasized that the ability to predict antigenic determinants from amino acid sequence data alone is potentially very useful, even though only a single determinant can be predicted with confidence for any given molecule. For example, many proteins whose antigenic structures are of interest are not available in quantities sufficient to allow conventional immunochemical studies to be carried out, as is the case with many of the proteins for which we listed predictions in the preceding section. Increasingly, amino acid sequence information for such proteins is being obtained by microchemical methods or by nucleotide sequence analysis, so that sufficient material for conventional immunochemical analysis is never available. However, once an antigenic determinant has been predicted, it should be possible to verify its existence by synthesizing the indicated region chemically and testing its activity in an appropriate immune assay, such as inhibition of cytotoxicity or precipitation inhibition. Furthermore, it should be possible to raise antisera against such synthetic determinants, as Arnon *et al.* have done for a bacteriophage (37). Ultimately, predicted antigenic determinants from proteins of pathogenic organisms might be useful in the production of synthetic vaccines.

1. Atassi, M. Z. (1975) *Immunochemistry* 12, 423–438.
2. Atassi, M. Z. & Lee, C. L. (1978) *Biochem. J.* 171, 429–434.
3. Reichlin, M. (1975) *Adv. Immunol.* 20, 71–123.
4. Jemmerson, R. & Margoliash, E. (1979) *J. Biol. Chem.* 254, 12706–12716.
5. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* 13, 222–224.
6. Levitt, M. (1976) *J. Mol. Biol.* 104, 59–107.
7. Maxfield, F. R. & Scheraga, H. A. (1976) *Biochemistry* 15, 5138–5153.
8. Rose, G. D. & Roy, S. (1980) *Proc. Natl. Acad. Sci. USA* 77, 4643–4647.
9. Nozaki, Y. & Tanford, C. (1971) *J. Biol. Chem.* 246, 2211–2217.
10. Kabat, E. A. (1968) *Structural Concepts in Immunology and Immunochemistry*, (Holt, Rinehart & Winston, New York), pp. 89–100.
11. Kelly, B. & Levy, J. G. (1971) *Biochemistry* 10, 1763.
12. Jemmerson R. & Margoliash, E. (1977) *Adv. Exp. Med. Biol.* 98, 119.
13. Eylar, E. H. (1977) *Adv. Exp. Med. Biol.* 98, 259.
14. Westall, F. C. & Thompson, M. (1978) *Immunochemistry* 15, 189.
15. Kazim, A. L. & Atassi, M. Z. (1977) *Biochem. J.* 167, 275.
16. Reichlin, M. & Noble, R. W. (1977) in *Immunochemistry of Proteins*, ed. Atassi, M. Z. (Plenum, New York), Vol. 2, p. 311.

17. Anderer, F. A. (1963) Z. *Naturforsch*. 18b, 1010.
18. Benjamini, E. (1977) in *Immunochemistry of Proteins*, ed. Atassi, M. Z. (Plenum, New York), Vol. 2, p. 265.
19. Milton, R. C. Del. & van Regenmortel, M. H. V. (1979) *Mol. Immunol*. 16, 179.
20. Kehoe, M. J. & Seide-Kehoe, R. (1979) in *Immunochemistry of Proteins*, ed. Atassi, M. Z. (Plenum, New York), Vol. 3, p. 87.
21. Hurrell, J. G. R., Smith, J. A. & Leach, S. J. (1978) *Immunochemistry* 15, 297–302.
22. Valenzuela, P., Gray, P., Quiroga, M., Zaldivar, J., Goodman, H. M. & Rutter, W. J. (1979) *Nature (London)* 280, 815–818.
23. Jou, W. M. Verhoeyen, M., Devos, R., Saman, E., Fang, R., Huylebroeck, D., Fiers, W., Threlfall, G., Barber, C. & Carey, N. (1980) *Cell* 19, 683–696.
24. Verhoeyen, M., Fang, R., Jou, W. M., Devos, R., Huylebroeck, D., Saman, E. & Fiers, W. (1980) *Nature (London)* 286, 771–776.
25. Porter, A. G., Barber, C., Carey, N. H., Hallewell, R. A., Threlfall, G. & Emtage, J. S. (1979) *Nature (London)* 282, 471–477.
26. Orr, H. T., Lopez de Castro, J. A., Lancet, D. & Strominger, J. L. (1979) *Biochemistry* 18, 5711–5720.
27. Derynck, R., Content, J., DeClercq, E., Volckaert, G., Tavernier, J., Devos, R. & Fiers, W. (1980) *Nature (London)* 285, 542–547.
28. Taniguchi, T., Mantei, N., Schwarzstein, M., Nagata, S., Muramatsu, M. & Weissman, C. (1980) *Nature (London)* 285, 547–549.
29. Goeddel, D. V., Yelverton, E., Ullrich A., Heyneker, H. L., Miozzari, G., Holmes, W., Seeburg, P. H., Dull, T., May, L., Stebbing, N., Crea, R., Maeda, S., McCandliss, R., Sloma, A., Tabor, J. M., Gross, M., Familletti, P. C. & Pestka, S. (1980) *Nature (London)* 287, 411–416.
30. Dallas, W. S. & Falkow, S. (1980) *Nature (London)* 288, 499–501.
31. So, M. & McCarthy, B. J. (1980) *Proc. Natl. Acad. Sci. USA* 77, 4011–4015.
32. Kurosky, A., Markel, D. E., Peterson, J. W. & Fitch, W. M. (1977) *Science* 195, 299–301.
33. Beachey, E. H., Seyer, J. M. & Kang, A. H. (1980) *J. Biol. Chem*. 255, 6284–6289.
34. Klapper, D. G., Goodfriend, L. & Capra, J. D. (1980) *Biochemistry* 19, 5729–5734.
35. Mole, L. E., Goodfriend, L., Lapkoff, C. B., Keoho, J. M. & Capra, J. D. (1975) *Biochemistry* 14, 1216–1220.
36. Garoff, H., Frischauf, A. M., Simons, K., Lehrach, H. & Delius, H. (1980) *Nature (London)* 288, 236–241.
37. Arnon, R., Sela, M., Parant, M. & Chedid, L. (1980) *Proc. Natl. Acad. Sci. USA* 77, 6769–6772.

Proc. Nat. Acad. Sci. USA
Vol. 68, No. 9, pp. 2293–2297, September 1971

Folding of Polypeptide Chains in Proteins: A Proposed Mechanism for Folding

PETER N. LEWIS, FRANK A. MOMANY, AND HAROLD A. SCHERAGA*

Department of Chemistry, Cornell University, Ithaca, N.Y. 14850

Contributed by Harold A. Scheraga, July 14, 1971

ABSTRACT **A mechanism is proposed for the folding of protein chains. On the basis of short-range interactions, certain aminoacid sequences have a high propensity to be, say, α-helical. However, these short helical (or other ordered) regions can be stabilized only by long-range interactions arising from the proximity of two such ordered regions. These regions are brought near each other by the directing influence of certain other aminoacid sequences that have a high probability of forming β-bends or variants thereof, also on the basis of short-range interactions. An analysis is made of the tendency of various amino acids to occur in β-bends, and it is possible to predict the regions of a chain in which a β-bend will occur with a high degree of reliability.**

In this series of papers, we will present a specific mechanism for the folding of a polypeptide chain into the native structure of a globular protein. In this presentation, we will attempt to demonstrate that specific backbone conformations such as the right-handed α-helix (α_R), the β-structure, and the β-bend, found to varying extents in the native structures of most globular proteins, are not only essential for the structural integrity of the protein but also are remnants of structures that play a key role in the folding process.

In this initial paper, we give a general description of the proposed mechanism, as well as some illustrative correlations between the aminoacid sequence and native structure of a protein that provide support for this mechanism. In subsequent papers in this series, we will discuss the energetics of the folding process.

PROPOSED MECHANISM

The protein molecule, under sufficiently denaturing conditions (or even, perhaps, directly after synthesis), behaves essentially as a random coil. Since the number of states accessible to the polypeptide chain in the random-coil condition is immense, it is reasonable to assume that (*a*) the folding of the chain into its most stable (native) conformation is *not* the result of a random event, and (*b*) a *specific* pathway exists for the folding process.

It was previously suggested (1) that one of the initial steps (which might be considered a nucleation step) during the folding process is the fortuitous meeting of two distant sections of the protein chain to form a stabilized pair of α-helices (or, for that matter, any other ordered structure), around which the rest of the polypeptide chain could fold. This idea developed from the demonstration (1) that, for most proteins, those portions of the chain that have a high helical probability in the denatured condition are found to be in the α_R conformation in the native structure. Further, it was shown (2) that, for the cytochrome *c* proteins of 27 species, the regions of high helical probability were, for the most part, conserved from species to species; this result is consistent not only with the proposed invariance of the native conformation of these proteins (3), but also with our proposal that these regions of high helical probability aid in directing the folding to the native structure.

While the above conclusion about one of the initial steps of the folding process seems warranted, it does not seem reasonable for the distant α-helical (or other ordered) conformations to rely on a *random* encounter to achieve a mutual stabilization (by means of long-range interactions) of the specific structures that have a propensity to be α-helical (because of short-range interactions) (1). Instead, it appears much more likely that two such distant helix-tending regions of the polypeptide chain are *directed* toward each other. The assumption of such a directing influence naturally introduces the proposition that certain regions of the chain function as "directing" sections. The role of these "directing" sections would be to provide the proper mutual orientation of distant (or near) ordered segments of the polypeptide chain so that the latter could interact with each other and, at the same time, serve as a substrate for interaction with still other chain segments. As an example, a β-bend or β-turn (defined in the next section) would "direct" the formation of an antiparallel β-structure which, in turn, might provide a surface for interaction (e.g., by means of hydrophobic bonds) with, and stabilization of, an α_R helix.

Our proposed mechanism for protein folding can be described as follows. A certain "directing" section promotes (in the manner described above) the formation of some small

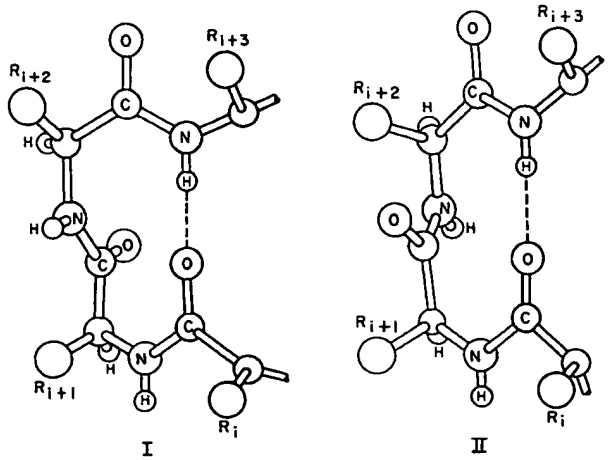


FIG. 1. Type I β-bend (*I*), with *any* L-residue at positions (i + 1) and (i + 2), and Type II β-bend (*II*), with only glycine (5) being possible at position (i + 2). Adapted from Fig. 7 of ref. 3.

* To whom requests for reprints should be addressed.

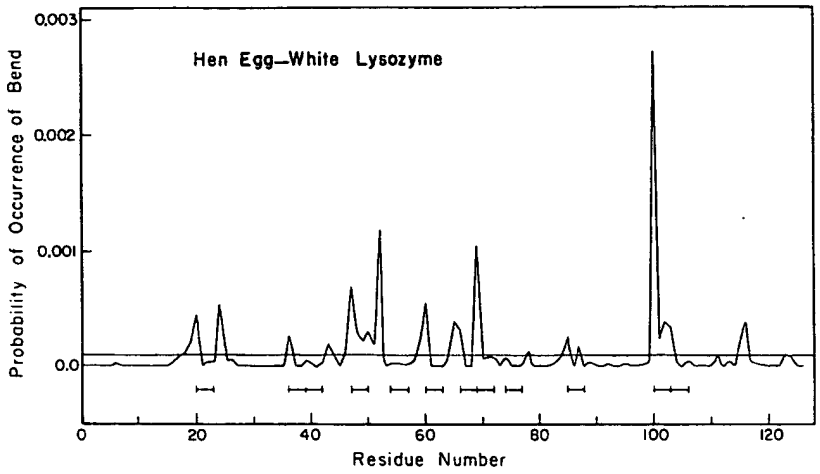


FIG. 2. Probability that a tetrapeptide bend begins at site j of the hen egg-white lysozyme chain. The horizontal line is an arbitrary cut-off probability. The horizontal bars indicate the positions of the observed bends (D. C. Phillips, personal communication), starting at $j = i$.

initial structure (e.g., two interacting α_R helical or other ordered structures). This ordered collection of aminoacid residues (backbone and side-chain groups) then serves as a substrate to direct other sections of the chain to either stabilize still other ordered structures (depending on the propensity of these additional segments to form such ordered structures) or simply to wrap around the original substrate†. If this mechanism is valid, then many of the distinctive features of the native structure (i.e., α_R helix or β-structure) would have been intimately involved in the folding pathway.

In the next section, the β-bend is shown to satisfy the conditions required for a "directing" section. It will be demonstrated that an analysis of the frequency of occurrence of particular aminoacid residues at the various loci of these bends in a sample of three proteins of known structure (namely, lysozyme, ribonuclease S, and α-chymotrypsin) provides sufficient information to enable us to predict, with a reliability of around 80%, the positions of bends that occur in the native conformations of other proteins.

THE β-BEND

In the previous section, it was argued that certain segments of the polypeptide chain are responsible for bringing distant portions of the chain into close proximity during the folding process. The simplest examples of such segments are the so-called β-bends, involving residues i to $(i + 3)$, as shown in Fig. 1. Venkatachalam (5) has considered the steric constraints in these bends for the formation of a hydrogen bond between the CO group of residue i and the NH group of residue $(i + 3)$; he showed that glycine must occur at position $(i + 2)$ for the bend in Fig. 1 (which is designated as a type-II bend) to involve such a hydrogen bond. However, for L-residues, there is no such restriction in the other bend shown in Fig. 1 (which is designated as a type-I bend). Both types of bends fulfill the requirements for being "directing" sections very nicely by (*a*) providing a possible 180° reversal of the chain direction, (*b*) having a high probability of occurrence, compared to a larger loop structure requiring long-range interactions, because the formation of the bend [and the i to $(i + 3)$ hydrogen bond] depends on only two pairs of dihedral angles (short-range interactions), and (*c*) involving only a small number (namely, four) of residues, thereby providing much conformational information in a small region.

Actually, the native structures of many proteins contain an abundance of β-bends, or β-like bends [distorted β-bends, which are similar to β-bends but lack the i to $(i + 3)$ hydrogen bond]. The distorted β-bend might arise by relaxation to satisfy the newly created interactions, once it has fulfilled its directing function during folding. In light of our earlier discussion of the possible importance of these bends for protein folding, it is of interest to consider the distribution of aminoacid residues at positions i, $(i + 1)$, $(i + 2)$, and $(i + 3)$ (see Fig. 1) in the actual bends found in the native conformations of some proteins. Toward this end, the coordinates of hen egg-white lysozyme‡, bovine ribonuclease S,(6) and the B and C chains of bovine α-chymotrypsin (7) were analyzed for bends. A bend was considered to exist if (*a*) the calculated $C^\alpha(i)$ to $C^\alpha(i + 3)$ distance was less than 7 Å (0.7 nm) and (*b*) the $(i + 1)$ or $(i + 2)$ residue was not in an α_R helix. The bends determined by the above criteria are given in Table 1§.

† It should be pointed out here that the notion that some initial structure (substrate) participates in the folding of a protein is not new and has been proposed by others (see, for example, ref. 4), although the emphasis placed in this paper on a "directing" section, such as a β-bend, has not to our knowledge been envisaged as an initial structure.

‡ D. C. Phillips, personal communication.

§ Criteria (*a*) and (*b*) in some cases are not sufficient to define the exact location of a bend. Molecular models of bovine ribonuclease S and bovine α-chymotrypsin, built in this laboratory, provided additional information concerning the existence and location of bends in these two proteins. The C-terminal (106–129) section of hen egg-white lysozyme is particularly difficult to analyze for bends by criteria (*a*) and (*b*), because of the presence of some helix in that part of the protein chain; therefore, for this region, no bends are given in Table 1, although it may be that some exist. It should be emphasized that the list of bends given in Table 1 may well be subject to revision and is presented here *only* to illustrate the possible role of "directing" sections. Further work is being performed, in this laboratory, to better characterize these bends.

TABLE 1. *β-bends and variants found in the native structures of hen egg-white lysozyme, bovine ribonuclease S, and bovine α-chymotrypsin (B and C chains)*

Hen egg-white lysozyme		Bovine ribonuclease S		Bovine α-chymotrypsin	
Number	Sequence	Number	Sequence	Number	Sequence
20–23	Tyr-Arg-Gly-Tyr	16–19	Ser-Thr-Ser-Ala	23–26	Val-Pro-Gly-Ser
36–39	Ser-Asn-Phe-Asn	36–39	Thr-Lys-Asp-Arg	27–30	Trp-Pro-Trp-Gln
39–42	Asn-Thr-Gln-Ala	65–68	Cys-Lys-Asn-Gly	35–38	Asp-Lys-Thr-Gly
47–50*	Thr-Asp-Gly-Ser	75–78	Ser-Tyr-Ser-Thr	48–51	Asn-Glu-Asn-Trp
54–57	Gly-Ile-Leu-Gln	87–90	Thr-Gly-Ser-Ser	56–59	Ala-His-Cys-Gly
60–63	Ser-Arg-Trp-Trp	91–94	Lys-Tyr-Pro-Asn	61–64	Thr-Thr-Ser-Asp
66–69	Asp-Gly-Arg-Thr	112–115	Gly-Asn-Pro-Tyr	72–75	Asp-Gln-Gly-Ser
69–72	Thr-Pro-Gly-Ser			91–94	Asn-Ser-Lys-Tyr
74–77	Asn-Leu-Cys-Asn			96–99	Ser-Leu-Thr-Ile
85–88	Ser-Ser-Asp-Ile			99–102	Ile-Asn-Asn-Asp
100–103*	Ser-Asp-Gly-Asp			108–111	Leu-Ser-Thr-Ala
103–106	Asp-Gly-Met-Asn			115–118	Ser-Gln-Thr-Val
				125–128	Ser-Ala-Ser-Asp
				131–134	Ala-Ala-Gly-Thr
				152–155	Pro-Asp-Arg-Leu
				172–175	Trp-Gly-Thr-Lys
				177–180	Lys-Asp-Ala-Met
				185–188	Ala-Ser-Gly-Val
				191–194	Cys-Met-Gly-Asp
				194–197	Asp-Ser-Gly-Gly
				203–206	Lys-Asn-Gly-Ala
				217–220	Ser-Ser-Thr-Cys
				221–224	Ser-Thr-Ser-Thr

* The $C^{\alpha}(i)$ to $C^{\alpha}(i + 3)$ distances for these two bends exceeded 7 Å by 0.1 Å and 0.4 Å for residues 47–50 and 100–103, respectively. Nevertheless, these bends were counted because, in both cases, each was the region of a significant chain reversal. See ref. 4 for stereo drawings of hen egg-white lysozyme.

TABLE 2. *Frequency of occurrence of amino acid residues in β-bends and variants found in hen egg-white lysozyme, bovine ribonuclease S, and bovine α-chymotrypsin (B and C chains)*

Amino acid	Total occurrence*	i	$i + 1$	$i + 2$	$i + 3$	$i \rightarrow (i + 3)$ (total†)	$\frac{i \rightarrow (i + 3) \text{ total}}{\text{(total occurrence)}}$
Ala	45	3	2	1	4	10	0.22
Asp	22	5	4	2	5	16	0.73
Cys	25	2	0	2	1	5	0.20
Glu	12	0	1	0	0	1	0.08
Phe	12	0	0	1	0	1	0.08
Gly	36	2	4	11	4	21	0.58
His	7	0	1	0	0	1	0.14
Ile	18	1	1	0	2	4	0.22
Lys	30	3	3	1	1	8	0.27
Leu	27	1	2	1	1	5	0.19
Met	8	0	1	1	1	3	0.38
Asn	36	4	4	3	4	15	0.42
Pro	13	1	4	1	0	6	0.46
Gln	19	0	2	1	2	5	0.26
Arg	18	0	2	2	1	5	0.28
Ser	51	11	6	6	5	28	0.55
Thr	39	5	4	6	4	19	0.49
Val	36	1	0	0	2	3	0.08
Trp	14	2	0	2	2	6	0.43
Tyr	13	1	2	1	2	6	0.46

* The numbers in this column represent the total occurrence of each residue in the three-protein sample.

† The $i \rightarrow (i + 3)$ totals for Asn, Thr, and Ile are each larger by 1 than their actual occurrence, because, for example, Thr simultaneously occupies positions i and $(i + 3)$ in bends 69–72 and 66–69, respectively, in lysozyme (see Table 1). Similarly, the total for Asp is larger by 2 than its actual occurrence.

TABLE 3. *Comparison of predicted and observed locations of bends in some globular proteins**

| Protein | | Location of ith bend position | | |

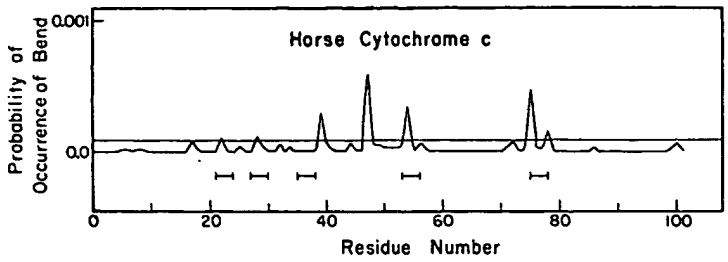


FIG. 3. Probability that a tetrapeptide bend begins at site j of the horse ferricytochrome c chain. The horizontal line is an arbitrary cut-off probability. The horizontal bars indicate the positions of the observed bends (3), starting at $j = i$. A bend (not shown here) occurs (3) between residues 42–46; this is not a β-bend (since it contains more than four residues) even though it results in a reversal of the chain.

tion as to the "directing" role of the bends in folding. This question is currently under investigation.

It should be pointed out that the β-bends discussed in this paper are thought to exist in specific sequences of amino acids. However, other β-bends (not considered here as necessarily likely to occur in globular proteins) can exist in *any* sequence of amino acids if the chain is short and constrained to form a ring, as in gramicidin S and oxytocin.

CONCLUSIONS

It is proposed that certain sections of a protein chain must play a role in bringing distant parts of the chain together to enable long-range interactions to stabilize those structures (i.e., α_R helix, β-structure, etc.) that have a propensity to form because of short-range interactions. The β-bend and its variants were shown to fulfill the requirements of a "directing" section. Further, it was shown that, to a good approximation, the distributions of aminoacid types in these bends are independent of each other; hence, the locations of a high percentage of the bends in proteins not included in the initial set of three can be predicted.

Since these bends in proteins are very localized, it would be interesting to determine (e.g., by NMR measurements) whether they occur in smaller structures, i.e., in isolated noncyclic tetra- or larger oligopeptides (with appropriate end groups).

In subsequent papers, we will consider the energetics of the β-bends and their variants.

P. N. L. was a National Research Council of Canada Postgraduate Fellow, 1971–1972. F. A. M. was a Special Fellow of the National Institute of General Medical Sciences, National Institutes of Health, 1968–1969. This work was supported by research grants from the National Science Foundation (GB-28469X and GB-17388), from the National Institute of General Medical Sciences of the National Institutes of Health, U.S. Public Health Service (GM-14312), from the Eli Lilly, Hoffmann-LaRoche, and Smith Kline and French Grants Committees, and from Walter and George Todd.

1. Lewis, P. N., N. Gō, M. Gō, D. Kotelchuck, and H. A. Scheraga, *Proc. Nat. Acad. Sci., USA*, 65, 810 (1970).
2. Lewis, P. N., and H. A. Scheraga, *Arch. Biochem. Biophys.*, 144, 576 (1971).
3. Dickerson, R. E., T. Takano, D. Eisenberg, O. Kallai, L. Samson, A. Cooper, and E. Margoliash, *J. Biol. Chem.*, 246, 1511 (1971).
4. Dickerson, R. E., and I. Geis, *The Structure and Action of Proteins* (Harper and Row, New York, 1969), p. 73.
5. Venkatachalam, C. M., *Biopolymers*, 6, 1425 (1968).
6. Wyckoff, H. W., D. Tsernoglou, A. W. Hanson, J. R. Knox, B. Lee, and F. M. Richards, *J. Biol. Chem.*, 245, 305 (1970).
7. Birktoft, J. J., B. W. Matthews, and D. M. Blow, *Biochem. Biophys. Res. Commun.*, 36, 131 (1969).
8. Arnone, A., C. J. Bier, F. A. Cotton, V. W. Day, E. E. Hazen, Jr., D. C. Richardson, J. S. Richardson, and, in part, A. Yonath, *J. Biol. Chem.*, 246, 2302 (1971).